

# CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training

Jianmin Bao<sup>\*1</sup>, Dong Chen<sup>2</sup>, Fang Wen<sup>2</sup>, Houqiang Li<sup>1</sup>, and Gang Hua<sup>2</sup>

<sup>1</sup>University of Science and Technology of China,  
jmbao@mail.ustc.edu.cn, lihq@ustc.edu.cn

<sup>2</sup>Microsoft Research Asia,  
{doch, fangwen, ganghua}@microsoft.com

## Abstract

We present variational generative adversarial networks, a general learning framework that combines a variational auto-encoder with a generative adversarial network, for synthesizing images of fine-grained categories, such as faces of a specific person or objects in a category. Our approach models an image as a composition of label and latent attributes in a probabilistic model. By varying the fine-grained category label fed to the resulting generative model, we can generate images in a specific category by randomly drawn values on a latent attribute vector. The novelty of our approach comes from two aspects. Firstly, we propose to adopt a cross entropy loss for the discriminative and classifier network, but a mean discrepancy objective for the generative network. This kind of asymmetric loss function makes the training of the GAN more stable. Secondly, we adopt an encoder network to learn the relationship between the latent space and the real image space, and use pairwise feature matching to keep the structure of generated images. We experiment with natural images of faces, flowers, and birds, and demonstrate that the proposed models are capable of generating realistic and diverse samples with fine-grained category labels. We further show that our models can be applied to other tasks, such as image inpainting, super-resolution, and data augmentation for training better face recognition models.

## 1. Introduction

Building effective generative models of natural images is one of the key problems in computer vision. It targets on generating diverse realistic images by varying some latent parameters according to the underneath natural image distributions. Therefore, a desired generative model

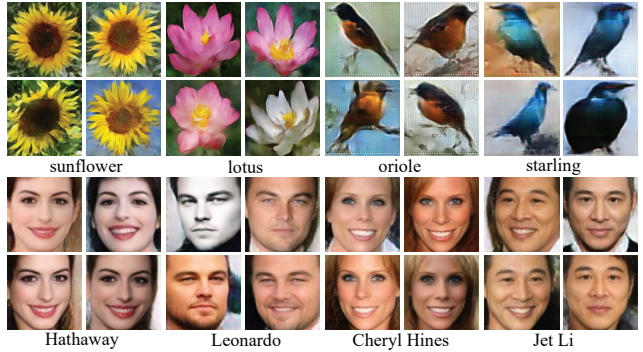


Figure 1. **Synthesized images** using our CVAE-GAN model at high resolution (128x128) for different classes. The generated samples are realistic and diverse within a class.

is necessitated to capture the underlying data distribution. This is often a very difficult task, since a collection of image samples may lie on a very complex manifold. Nevertheless, recent advances in deep convolutional neural network nourish a series of deep generative models [14, 12, 8, 31, 29, 34, 15, 4, 33, 6] that made tremendous progress, largely due to the capability of deep networks in learning representations.

Building on top of the success of these recent works, we are interested in going one step further to generate images of fine-grained object categories. For example, we would like to be able to synthesize images for a specific identity (Figure 1), or produce a new image of a specified specie of flowers or birds, and so on. Inspired by CVAE [34] and VAE/GAN [15], we propose a general learning framework that combines a variational auto-encoder with a generative adversarial network under a conditioned generative process to tackle this problem.

However, we found this naive combination insufficient in practice. The results from VAE is usually blurry. The discriminator can easily classify them as “fake”. Even though they sometimes looks remarkably good for face images.

<sup>\*</sup>This work was done when Jianmin Bao was an intern at MSR Asia.

The gradient vanishing problem still exists. So the generated images are very similar to the results using VAE alone.

In this paper, we propose a new objective for the generator. Instead of using the same cross entropy loss as the discriminator network, the new objective requires the generator to generate data that minimize the  $\ell_2$  distance of average feature to the real data. For multi-class image generation, the generated samples of one category also need to match the average feature of real data of that category. Since the feature distance and the separability are positively correlated. It solved the gradient vanishing problem to a certain extent. Besides, this kind of asymmetric loss function can partially help prevent mode collapse problem that all outputs moving toward a single point, it makes the training of GAN more stable.

Although using mean feature matching will reduce the chance of mode collapse, it did not completely solve this problem. Once mode collapse occurred, the gradient descent is not able to separate the identical outputs. To keep the diversity of generated samples, we take advantage of the combination of VAE and GAN. We use an encoder network to map the real image to the latent vector. Then the generator is required to reconstruct the raw pixels and match the feature of original images with a given latent vector. In this way, we explicitly set up the relationship between the latent space and real image space. Because the existence of these anchor points, the generator are enforced to emit diverse samples. Moreover, the pixel reconstruction loss is also helpful to keep the structure, such as straight line and face structure in the image.

As shown in Figure 2 (g), our framework consists of four parts: 1) the encoder network  $E$ , which maps the data sample  $x$  to a latent representation  $z$ . 2) The generative network  $G$ , which generates image  $x'$  given a latent vector. 3) The discriminative network  $D$ , which distinguishes real/fake images. 4) The classifier network, which measures the class probability of the data. We named our approach as CVAE-GAN. These four parts are seamlessly cascaded together, and the whole pipeline is trained end-to-end.

Once the CVAE-GAN is trained, it can be used in different applications, *e.g.*, image generation, image inpainting, and attributes morphing. Our approach estimates a good representation of the input image, and the generated image appears to be more realistic. We show that it outperforms CVAE and CGAN and other state-of-the-art methods. Comparing with GAN, the proposed framework is much easier to train and converges faster and more stable in the training stage. In the experiment, we further show that the images synthesized from our models can be applied to other tasks, such as data augmentation for training better face recognition models.

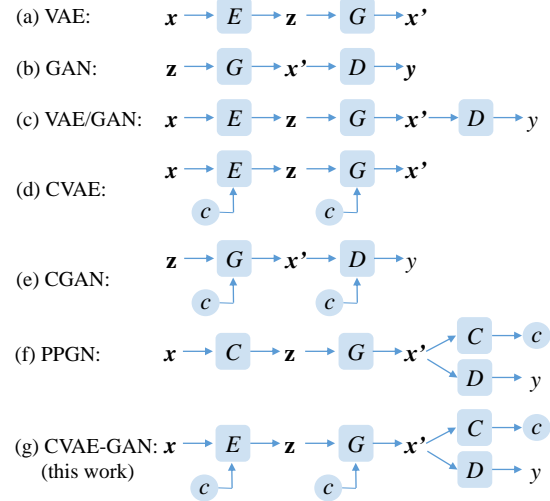


Figure 2. Illustration of the structure of VAE [12, 31], GAN [8], VAE/GAN [15], CVAE [34], CGAN [18], PPGN [23] and the proposed CVAE-GAN. Where  $x$  and  $x'$  are input and generated image.  $E, G, C, D$  are encoder, generative, classification, and discriminative network, respectively.  $z$  is the latent vector.  $y$  is a binary output which represents real/synthesized image.  $c$  is the condition, such as attribute or class label.

## 2. Related work

Conventional wisdom and early research of generative models, including Principle Component Analysis (PCA) [40], Independent Component Analysis (ICA) [10], Gaussian Mixture Model (GMM) [46, 27, 37], all assume simple formation of the data. They have difficulty in modeling complex patterns of irregular distributions. Later works, such as Hidden Markov Model (HMM) [35], Markov Random Field (MRF) [19], restricted Boltzmann machines (RBMs) [9, 32], discriminatively trained generative models [39], also limit their results on texture patches, digital number or well aligned faces, due to lack of effective feature representations.

Recent development of deep generative models, [14, 12, 8, 31, 29, 15, 4, 33, 6] has caught the attention of a lot of researchers. Since deep hierarchical architectures allow them to capture complex structures in the data, all these methods show promising results of generating natural images that are far more realistic than conventional generative models. Among them, there are three main themes: Variational Auto-encoder (VAE) [12, 31], Generative Adversarial Network (GAN) [8, 29, 33], and Autoregression [14].

VAE [12, 31] pairs a differentiable encoder network with a decoder/generative network. A disadvantage of VAE is that, because of the injected noise and imperfect element-wise measures such as the squared error, the generated samples are often blurry.

Generative Adversarial Network (GAN) [8, 29, 33] is another popular generative model. It simultaneously trains

two models: a generative model to synthesize samples, and a discriminative model to differentiate between natural and synthesized samples. However, the GAN model is hard to converge in the training stage and the samples generated from GAN are often far from natural. Recently, many works tried to improve the quality of the generated samples. For example, Wasserstein GAN (WGAN) [2] used Earth Mover Distance as an objective for training GANs, and McGAN [20] used mean and covariance feature matching. But they need to limit the range of the parameters of discriminator which will decrease the discriminative power. Loss-Sensitive GAN [28] learned a loss function which can quantify the quality of generated samples and used this loss function to generate high-quality images. There are also methods which tried to combine GAN and VAE, *e.g.*, VAE/GAN [15] and adversarial autoencoders [17]. They are very related to and partly inspired our work.

VAEs and GANs can also be trained to conduct conditional generation, *e.g.*, CVAE [34] and CGAN [18]. By introducing additional conditionality, they can handle probabilistic one-to-many mapping problem. Recently there are many interesting works based on CVAE and CGAN, including conditional face generation [7], Attribute2Image [47], text to image synthesis [30], forecasting from static images [42], and conditional image synthesis [25]. All of them obtained amazing results.

Generative ConvNet [44], which demonstrated that a generative model can be derived from the commonly used discriminative ConvNet. Dosovitskiy et al. [5] and Nguyen et al. [22] introduces a method which generates high quality images from features extracted from a trained classification model. PPGN [23] performs excellently in generating samples by using a gradient ascent and prior to the latent space of a generator.

Autoregression [14] follows a different idea. It uses autoregressive connections to model images pixel by pixel. Its two variants, PixelRNN [41] and PixelCNN [26], also produce excellent samples.

Our model differs from all these models. As illustrated in Figure 2, we compare the structure of the proposed CVAE-GAN with all these models. Besides the difference in the structure, more importantly, we take advantages of both statistic and pairwise feature matching to make the training process converge faster and more stable.

### 3. Our formulation: the CVAE-GAN

In this section, we will introduce the proposed CVAE-GAN networks. As shown in Figure 3, our proposed method contains four parts: 1) the encoder network  $E$ ; 2) the generative network  $G$ ; 3) the discriminative network  $D$ ; 4) and the classification network  $C$ .

The function of network  $E$  and  $G$  is the same as that in conditional variational auto-encoder (CVAE) [34]. The en-

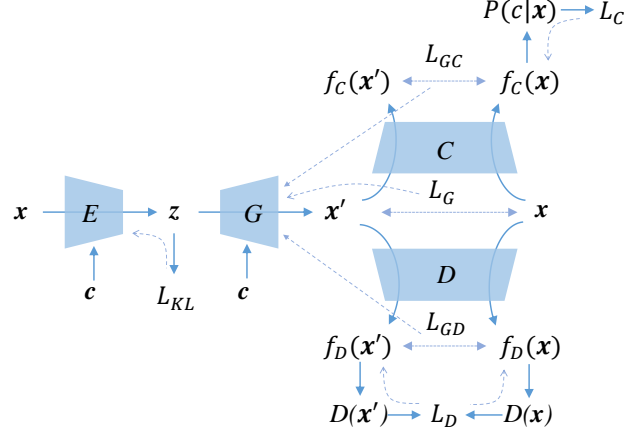


Figure 3. Illustration of our network structure. Our model contains four parts: 1) The encoder network  $E$ ; 2) The generative network  $G$ ; 3) The classification network  $C$ ; 4) The discriminative network  $D$ . Please refer to Section 3 for the detail.

coder network  $E$  maps the data sample  $x$  to a latent representation  $z$  through a learned distribution  $P(z|x, c)$ , where  $c$  is the category of the data. The generative network  $G$  generates image  $x'$  by sampling from a learned distribution  $P(x|z, c)$ . The function of network  $G$  and  $D$  is the same as that in the generative adversarial network (GAN) [8]. The network  $G$  tries to learn the real data distribution by the gradients given by the discriminative network  $D$  which learns to distinguish between “real” and “fake” samples. The function of network  $C$  is to measure the posterior  $P(c|x)$ .

However, the naive combination of VAE and GAN is insufficient. Recent work [1] shows that if the original  $KL$  Divergence loss is adopted, training of GAN will suffer from a gradient vanishing problem of the network  $G$ . Therefore, we only keep the training process of network  $E$ ,  $D$ , and  $C$  as the same as the original VAE [12] and GAN [8], and propose a new mean feature matching objective for the generative network  $G$  to improve the stability of the original GAN.

Even with the mean feature matching objective, there is still some risks to cause the mode collapse. So we use the encoder network  $E$  and the generative network  $G$  to obtain a mapping from real samples  $x$  to the synthesized samples  $x'$ , by using the pixel-wise  $\ell_2$  loss and pair-wise feature matching, the generative model is enforced to emit diverse samples and generate structure-preserving samples.

In the following sections, we first describe the method of mean feature matching based GAN (Section 3.1). Then we show that the mean feature matching can also be used in conditional image generation tasks (Section 3.2). After that, we introduce pair-wise feature matching by using an additional encoder network (Section 3.3). Finally, we analyse the objective of the proposed method and provide the implementation detail in the training pipeline (Section 3.4).

### 3.1. Mean feature matching based GAN

In traditional GANs, the generator  $G$  and a discriminator  $D$  compete in a two-player minimax game. The discriminator tries to distinguish real training data from synthesized images, and the generator tries to fool the discriminator. Concretely, the network  $D$  tries to minimize the loss function

$$\mathcal{L}_D = -\mathbb{E}_{\mathbf{x} \sim P_r} [\log D(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim P_z} [\log(1 - D(G(\mathbf{z})))] \quad (1)$$

while network  $G$  tries to minimize

$$\mathcal{L}'_{GD} = -\mathbb{E}_{\mathbf{z} \sim P_z} [\log D(G(\mathbf{z}))].$$

However, in practice, the distributions of “real” and “fake” images may not with each other, especially at the early stage of the training process. The discriminative network  $D$  could perfectly separate them. That is, we always have  $D(\mathbf{x}) \rightarrow 1$  and  $D(\mathbf{x}') \rightarrow 0$ , where  $\mathbf{x}' = G(\mathbf{z})$  is the generated image. Therefore, when updating network  $G$ , the gradient  $\partial \mathcal{L}'_{GD} / \partial \mathbf{x}' \rightarrow 0$ . The network  $G$  will easily get trapped in a local minimal, since  $G$  is not a convex function. Recent works [1, 2] also theoretically shown that training of GAN is often confronted by the vanishing gradient of  $G$ .

To address this problem, we proposed to use a mean feature matching objective for the generator. The objective requires the center of the features of the synthesized samples to match the center of the feature of the real samples. Let  $f_D(\mathbf{x})$  denote features on an intermediate layer of the discriminator, then  $G$  tries to minimize the loss function

$$\mathcal{L}_{GD} = \frac{1}{2} \|\mathbb{E}_{\mathbf{x} \sim P_r} f_D(\mathbf{x}) - \mathbb{E}_{\mathbf{z} \sim P_z} f_D(G(\mathbf{z}))\|_2^2 \quad (2)$$

In our experiment, for simplicity, we chose the input of the last Fully Connected (FC) layer of the network  $D$  as the feature  $f_D$ . Combining the features of multiple layers could marginally improve the converging speed. In the training stage, we estimate the mean feature using the data in a minibatch. And we also use moving average with the history to make it more stable.

Therefore, in the training stage, we update network  $D$  using Eq. 1, and update network  $G$  using Eq. 2. Using this asymmetrical loss for training GAN has the following three advantages: 1) since Eq. 2 increases with the separability, the  $\ell_2$  loss on feature center solves the gradient vanishing problem; 2) when the generated images are good enough, the mean feature matching loss become zero, it makes the training more stable; 3) comparing with WGAN [2], we do not need to clip the parameters. The discriminative power of the network  $D$  can be kept.

### 3.2. Mean feature matching for conditional image generation

In this section, we will introduce mean feature matching for conditional image generation. Suppose we have a

set of data belonging to  $K$  categories. We use the network  $C$  to measure whether an image belongs to a specific fine-grained category. Here we use a standard method for classification. The network  $C$  takes in  $\mathbf{x}$  as input and outputs a  $K$ -dimensional vector, which then turns into class probabilities using a softmax function. The output of each entry represents the posterior probability  $P(c|\mathbf{x})$ . In the training stage, the network  $C$  tries to minimize the softmax loss

$$\mathcal{L}_C = -\mathbb{E}_{\mathbf{x} \sim P_r} [\log P(c|\mathbf{x})]. \quad (3)$$

While for the network  $G$ , if we still use the similar softmax loss function as Eqn. 3. It will suffer from the same gradient vanishing problem as described in Section 3.1.

Therefore, we proposed to use the mean feature matching objective for the generative network  $G$ . Let  $f_C(\mathbf{x})$  denote features on an intermediate layer of the classification, then  $G$  tries to minimize:

$$\mathcal{L}_{GC} = \frac{1}{2} \sum_c \|\mathbb{E}_{\mathbf{x} \sim P_r} f_C(\mathbf{x}) - \mathbb{E}_{\mathbf{z} \sim P_z} f_C(G(\mathbf{z}, c))\|_2^2. \quad (4)$$

Here, we choose the input of the last FC layer of network  $C$  as the feature for simplicity. We also try to combine features of multiple layers, it only marginally improves the ability of identity preservation of network  $G$ . Since there are only a few samples belongs to the same category in a minibatch, it is necessary to use moving average of features for both real and generated samples.

### 3.3. Pairwise feature matching

Although, using mean feature matching could prevent all outputs moving toward a single point, such that reducing the chance of mode collapse, it does not completely solve this problem. Despite that the generated samples and real samples have the same feature center, they may have different distributions. Once mode collapse occurred, the generative network outputs the same images for different latent vectors, thus the gradient descent will not be able to separate these identical outputs.

In order to generate diverse samples, DCGAN [29] uses Batch Normalization, McGAN [20] uses both mean and covariance feature statistics, Salimans *et al.* [33] use minibatch discrimination. They are all based on using multiple generated examples. Different from these methods, we add an encoder network  $E$  to obtain a mapping from the real image  $\mathbf{x}$  to the latent space  $\mathbf{z}$ . Therefore, we explicitly set up the relationship between the latent space and real image space.

Similar to VAE, for each sample, the encoder network outputs the mean and covariance of the latent vector, i.e.,  $\mu$  and  $\epsilon$ . We use the  $KL$  loss to reduces the gap between the prior  $P(\mathbf{z})$  and the proposal distributions, i.e.,

$$\mathcal{L}_{KL} = \frac{1}{2} (\mu^T \mu + \text{sum}(\exp(\epsilon) - \epsilon - 1)). \quad (5)$$

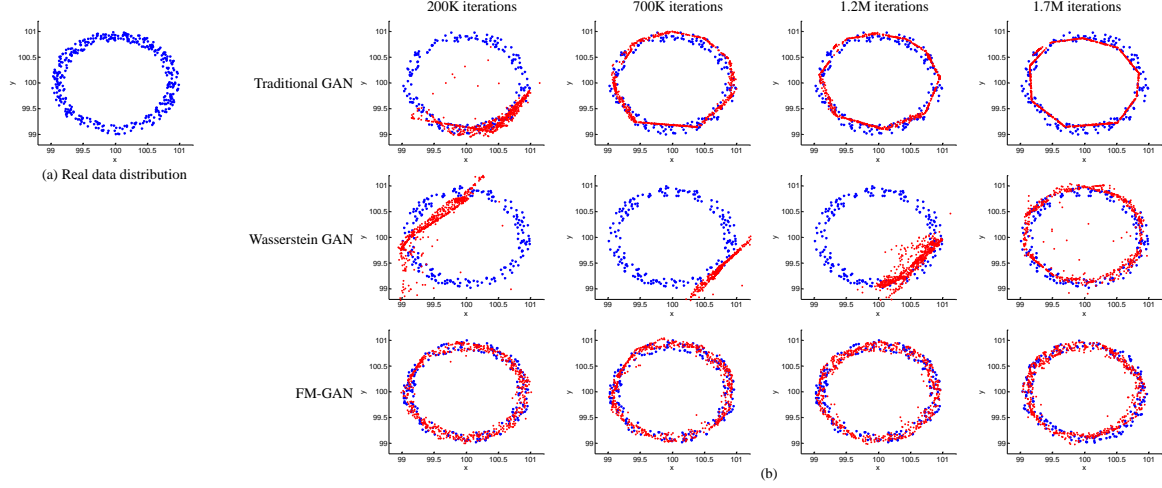


Figure 4. Result on a toy example for different generative models. The blue dots are the real points, the red dots are the generated points. a) the real data distribution which is like a “ring”; b) The generated points by traditional GAN, WGAN and mean feature matching GAN at different iterations. The traditional GAN suffers from the mode collapse problem, the generated samples lie in a thin circle, and cannot fit the whole real data distribution. WGAN successfully learns real data distribution, but the convergence speed is much slower than our method.

Then, we can sample the latent vector  $\mathbf{z} = \mu + \mathbf{r} \odot \exp(\epsilon)$ , where  $\mathbf{r} \sim N(\mathbf{0}, \mathbf{I})$  is a random vector and  $\odot$  represents the element-wise multiplication. After obtaining a mapping from  $\mathbf{x}$  to  $\mathbf{z}$ , we can obtain the generated image  $\mathbf{x}'$  with the network  $G$ . Then, we can add a  $\ell_2$  reconstruction loss and pair-wise feature matching loss between  $\mathbf{x}$  and  $\mathbf{x}'$ ,

$$\mathcal{L}_G = \frac{1}{2}(\|\mathbf{x} - \mathbf{x}'\|_2^2 + \|f_D(\mathbf{x}) - f_D(\mathbf{x}')\|_2^2 + \|f_C(\mathbf{x}) - f_C(\mathbf{x}')\|_2^2), \quad (6)$$

where,  $f_D$  and  $f_C$  are the features of an intermediate layer of the discriminative network  $D$  and classification network  $C$ , respectively.

### 3.4. Objective of CVAE-GAN

Therefore, to sum up, the goal of our approach is to minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{KL} + \mathcal{L}_G + \mathcal{L}_{GD} + \mathcal{L}_{GC} + \mathcal{L}_D + \mathcal{L}_C, \quad (7)$$

where the exact forms of each of the terms are presented in Eqns. 1~10. Every term of the above formula is meaningful.  $\mathcal{L}_{KL}$  is only related to the encoder network  $E$ . It represents whether the distribution of the latent vector is under expectation.  $\mathcal{L}_G$ ,  $\mathcal{L}_{GD}$  and  $\mathcal{L}_{GC}$  are related to the generative network  $G$ . They represent whether the synthesized image is similar to the input training sample, the real image, and other samples within the same category, respectively.  $\mathcal{L}_C$  is related to the classification network  $C$ , which represents the capability of the network to classify images from different categories, and  $\mathcal{L}_D$  is related to the discriminative network, which represents how good is the network

to distinguish between real/synthesized images. All these objectives are complementary to each other, and ultimately enable our algorithm to obtain superior results. The whole training procedure is described in Algorithm 1

## 4. Analysis with a toy example

In this section, we present and demonstrate the benefits of mean feature matching based GAN with a toy example. We assume that we have a real data distribution which is a “ring” as shown in Figure 4(a). The center of the ring is set to (100, 100), such that it is far from the generated distribution at the beginning. We compare the traditional GAN, WGAN, and the mean feature matching based GAN introduced in Section 3.1 to learn the real data distribution.

The three compared models share the same setting, the generator  $G$  is a MLP with 3 hidden layers with 32, 64, 64 units, respectively. The discriminator  $D$  is also a MLP with 3 hidden layers with 32, 64, 64 units, respectively. We used RMSProp and fixed learning rate 0.00005 for all methods. We trained each model for 2M iterations until they are all converged. The generated samples of each model at different iterations are plotted in Figure 4. From the results we can observe that:

1. For traditional GAN (first row in Figure 4(b)), the generated samples only lie in a limited area of the real data distribution, which is known as the mode collapse problem. This problem always exist during the training process.
2. For WGAN (second row in Figure 4(b)), it cannot learn the real data distribution at early iterations, we think this problem is caused by the clamping weights



**Algorithm 1** The training pipeline of the proposed CVAE-GAN algorithm.

**Require:**  $m$ , the batch size.  $n$ , class number.  $\theta_E$ , initial  $E$  network parameters.  $\theta_G$ , initial  $G$  network parameters.  $\theta_D$ , initial  $D$  network parameters.  $\theta_C$ , initial  $C$  network parameters.

```

1: while  $\theta_G$  has not converged do
2:   Sample  $\{x_r, c\} \sim P_r$  a batch from the real data;
3:    $\mathcal{L}_C \leftarrow -\log(P(c|x_r))$ 
4:    $z \leftarrow E(x_r, c)$ 
5:   Sample  $z_p \sim P_z$  a batch of prior samples;
6:    $\mathcal{L}_{KL} \leftarrow KL(q(z|x_r, c)||P_z)$ 
7:    $x_f \leftarrow G(z, c)$ 
8:    $x_p \leftarrow G(z_p, c)$ 
9:    $\mathcal{L}_D \leftarrow -(\log(D(x_r)) + \log(1 - D(x_f)) + \log(1 - D(x_p)))$ 
10:  Calculate  $x_r$  feature center  $\frac{1}{m} \sum_i^m f_D(x_r)$  and  $x_p$ 
    feature center  $\frac{1}{m} \sum_i^m f_D(x_p)$ ;
11:   $\mathcal{L}_{GD} \leftarrow \frac{1}{2} \|\frac{1}{m} \sum_i^m f_D(x_r) - \frac{1}{m} \sum_i^m f_D(x_p)\|_2^2$ 
12:  Calculate each class  $c_i$  feature center  $f_C^{c_i}(x_r)$  for  $x_r$ 
    and  $f_C^{c_i}(x_p)$  for  $x_p$  using moving average method;
13:   $\mathcal{L}_{GC} \leftarrow \frac{1}{2} \sum_{c_i} \|f_C^{c_i}(x_r) - f_C^{c_i}(x_p)\|_2^2$ 
14:   $\mathcal{L}_G \leftarrow \frac{1}{2} (\|x_r - x_f\|_2^2 + \|f_D(x_r) - f_D(x_f)\|_2^2 + \|f_C(x_r) - f_C(x_f)\|_2^2)$ 
15:   $\theta_C \leftarrow \theta_C + \nabla_{\theta_C}(\mathcal{L}_C)$ 
16:   $\theta_D \leftarrow \theta_D + \nabla_{\theta_D}(\mathcal{L}_D)$ 
17:   $\theta_G \leftarrow \theta_G + \nabla_{\theta_G}(\mathcal{L}_G + \mathcal{L}_{GD} + \mathcal{L}_{GC})$ 
18:   $\theta_E \leftarrow \theta_E + \nabla_{\theta_E}(\mathcal{L}_G + \mathcal{L}_{KL})$ 
19: end while

```

trick. Which influence  $D$ 's capability in distinguishing real/fake samples. We also tried to vary the clamp values to accelerate the training process, and find that if the value is too small, it will cause the gradient vanishing problem. If too large, the network will diverge.

- Third row shows the result of the proposed feature matching based GAN. It correctly learns the real data distribution the fastest.

## 5. Experiments

In this section, we will experimentally validate the proposed method. We evaluated our model on three datasets: the FaceScrub [21], the CUB-200 [43], and 102 Category Flower [24] datasets. These three datasets contain three completely different objects, which are human faces, birds, and flowers, respectively. It tests the generalization ability of our model.

The sizes of input and synthesized images are  $128 \times 128$  for all experiments. For FaceScrub dataset, we first detect

the face region with the JDA face detector [3], and then locate five facial landmarks (two eyes, nose tip and two mouth corners) with SDM [45]. After that, we use similarity transformation based on the facial landmarks to align faces to a canonical position. Finally, we crop a  $128 \times 128$  face region centered around the nose tip. For CUB-200 dataset, we just use the original images from the dataset. For 102 Category Flower dataset, we tightly crop a rectangle region based on the ground-truth mask which contains the flower, and then resize it into  $128 \times 128$ .

In our experiments, the encoder network  $E$  is a GoogleNet [36], The category information and the image is merged at the last FC layer of the encoder network  $E$ . The generative network  $G$  consists of 2 fully-connected layers, followed by 6 deconv layers with 2-by-2 upsampling. The convolution layers have 256, 256, 128, 92, 64 and 3 channels with filter size of  $3 \times 3$ ,  $3 \times 3$ ,  $5 \times 5$ ,  $5 \times 5$ ,  $5 \times 5$ ,  $5 \times 5$ . For the discriminative network we use the same discriminative network as the DCGAN [29]. For the classification network, we use a Alexnet [13] structure, and change the input to  $128 \times 128$ . We fixed the latent vector dimension to be 256 and found this configuration sufficient for generating images. The batch normalization layer [11] is also applied after each convolution layer. The model is implemented using deep learning toolbox Torch.

### 5.1. Visualization comparison with other models

In this experiment, we compare the proposed mean feature matching based CGAN introduced in Section 3.2(FM-CGAN), and CVAE-GAN model with other generative models for image synthesis of fine-grained image.

In order to fairly compare each method, we use the same network structure for all methods. For CVAE, we used the same convolution architecture from the encoder network  $E$  and the generative network  $G$ . For CGAN, we use the same generative network  $G$  and the same discriminative network  $D$  as in our CVAE-GAN method. All methods are trained with the same training data. And in the testing stage, the network architectures are the same. All three methods only use the network  $G$  to generate images. Therefore, although our approach has more parameters in the training stage, we believe that this comparison is fair.

We conduct the experiments on three datasets: FaceScrub, CUB-200 and 102 Category Flower dataset. We performed the task of category conditioned image generation for all methods. For each dataset, all methods are trained with all the data in that dataset. In the test stage, we first randomly chose a category  $c$ , and then randomly generate samples of that category by sampling the latent vector  $z \sim N(\mathbf{0}, \mathbf{I})$ . For evaluation, we visualized the samples generated from all the methods.

The comparison results are presented in Figure 5. All pictures are randomly selected without any personal bias.

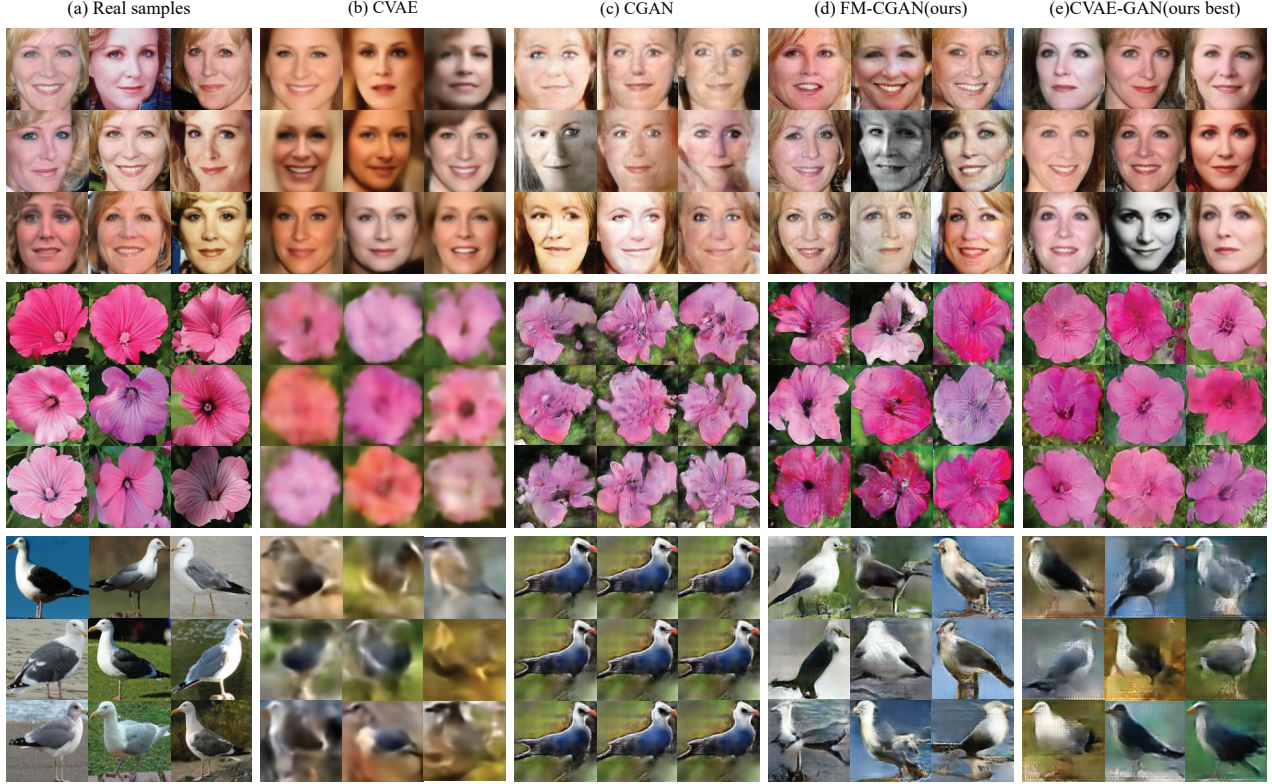


Figure 5. Comparison of random generated samples from different methods on FaceScrub [21], CUB-200 [43] and 102 Category Flower datasets [24]. We encourage readers to zoom in the figure to find the details of synthesised images. a) 9 random real image of one category. b) results of CVAE, which is blurry and cannot preserve the category identity, c) results of traditional CGAN, it loses diversity and structure info, d) results of our mean feature matching CGAN, it shows diverse results, but also losing structure info. e) results of our CVAE-GAN, which shows realistic, diversity and category-keeping results.

	Real data	CVAE	CGAN	FM-CGAN	CVAE-GAN
Top-1 acc	99.61%	8.09%	61.97%	79.76%	97.78%
Realisticity	20.85	10.29	15.79	19.40	19.03

Table 1. Quantitative result of generated image quality of different methods. Please refer to Section 5.2 for the detail.

We can observe that the image generated by CVAE is often blurry. We also notice that CVAE may not keep the identity information in the image, some faces generated from CVAE do not look similar to the person of that category in the FaceScrub dataset. For traditional CGAN, the variation within a category is very small, which is resulted from the mode collapse. For FM-CGAN, we can observe clear images with well preserved identity, but some images lose the structure of an object, such as the shape of the face.

On the other hand, images generated by the proposed CVAE-GAN models look realistic and clear, and are non-trivially different from each other, especially for view-point and background color. Our model is also able to keep the identity information. It shows the strength of the proposed CVAE-GAN method.

## 5.2. Quantitative comparison

Evaluating the quality of synthesized image is challenging due to the variety of probabilistic criteria [38]. We attempt to measure the generative model from three aspects: discriminability, diversity and realisticity.

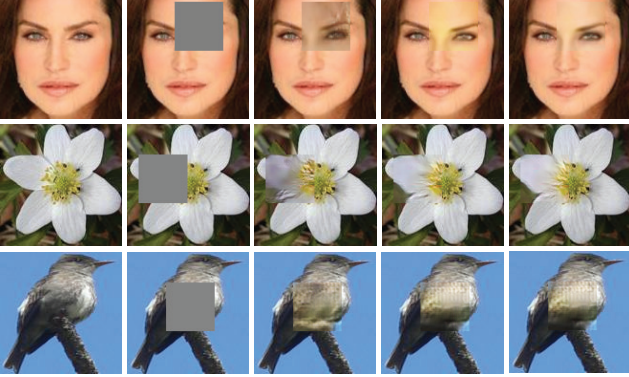
We use face images for this experiment. First of all, we randomly generated 53k samples (100 for each class) from CVAE, CGAN, FM-CGAN, and CVAE-GAN models for evaluation.

To measure the discriminability, we use a pre-trained face classification network on the real data. Here we use the Inception network [36]. With this trained model, we evaluate top-1 accuracy of the generated samples from each method. The results are shown in Table 1. As we can see, our model achieves the best top-1 accuracy with a big gap to other generative models. It demonstrates the effectiveness of the proposed method.

Follow the method in [33], we use the *Inception Score* to evaluate the realisticity and diversity of generated samples. We trained a classification model on the CASIA [48] datasets, and adopted  $\exp(\mathbb{E}_{\mathbf{z}} KL(p(y|\mathbf{x})||p(y)))$  as the metric to measure the realisticity and diversity of the gener-



Figure 6. Results of attributes morphing.



a) Original images b) Masked images c) CVAE-GAN-1 d) CVAE-GAN-5 e) CVAE-GAN-10  
Figure 7. Result of image inpainting by our proposed model CVAE-GAN-1  $\sim$  10 shows the results of iteration 1  $\sim$  10.

active models, where  $p(y|x)$  represents the posterior probability of each class of the generated samples. Images that contain meaningful objects should have a conditional label distribution  $p(y|x)$  with low entropy. Moreover, if the model generate diverse images, the marginal  $p(y) = \int p(y|G(z))dz$  should have high entropy. The larger score means the generator can produce more realistic and diverse images. As shown in Table 1. Our proposed CVAE-GAN and FM-CGAN achieve better score than other models, which is even very close to the real data.

### 5.3. Attributes morphing

In this part, we will validate that the attribute in the generated images will be continuously changed with the latent vector. We call this phenomenon attribute morphing. We also test our model on FaceScrub, CUB-200 and 102 Category Flower datasets. We first select a pair of images  $x_1$  and  $x_2$  in the same category, and then extract the latent vector  $z_1$  and  $z_2$  using the encoder network  $E$ . Finally, we can obtain a series of latent vectors  $z$  by linear interpolation, i.e.,  $z = \alpha z_1 + (1 - \alpha)z_2, \alpha \in [0, 1]$ . Figure 6 shows the results of attribute morphing. In each row, the attribute, such as pose, emotion, color, or flower number, gradually change from left to right.

### 5.4. Image inpainting

In this part, we show that our model can also be applied for image inpainting. We first randomly corrupt a  $50 \times 50$  patch of an original  $128 \times 128$  image  $x$  (Fig.7b), and then

Method	Training Data	Accuracy
no data augmentation	80K	91.87%
existing identities augmentation	80K + 100K	92.77%
5k new identities augmentation	80K + 500K	92.98%

Table 2. Results of face data augmentation.

feed it to the  $E$  network to obtain a latent vector  $z$ , then we can synthesize an image  $x'$  by  $G(z, c)$  where  $c$  is the class label, then we update the image by the following equation, i.e.,

$$x = M \odot x' + (1 - M) \odot x, \quad (8)$$

where  $M$  is the binary mask for the corrupted patch,  $\odot$  denotes the element-wise product. So  $(1 - M) \odot x$  is the uncorrupted area in the original image. The inpainting result is shown in Figure 7 (c). We should emphasize that all input images are download from website, none of them belong to the training data. Of course, we can iteratively feed the result images to the model to obtain a better results, as shown in Figure 7 (d,e).

### 5.5. CVAE-GAN for data augmentation

We further show that the images synthesized from our model can be used for data augmentation for training better face recognition model. We use the FaceScrub dataset as training data, and test on the LFW [16] dataset.

FaceScrub dataset only contains 530 persons, which is insufficient to train a good recognition model. Therefore, our goal is to verify that the use of additional generated training data can improve the face recognition accuracy. A larger dataset may be used for the training. We leave this for future study. We adopt the GoogleNet with softmax loss as our network.

We experimented two data augmentation strategies: 1) generate more images for existing identities in the training datasets; 2) generating new identities by mixing of different identities. We test this two kind of data augmentation methods. For 1), we randomly generate about 200 images per person, totally 100k images. For 2), we create 5k new identities by randomly mixing the label of three different existing identities, and generate 100 images for each new identity. For both strategies, the generated images are combined with the Facescrub dataset to train a face recognition model.

In the testing stage, we directly use the cosine similarity of features to measure the similarity between two faces. In Table 2, we compare the face recognition accuracy on the LFW dataset with and without additional synthesized faces. With new identities data augmentation of new identities, we achieve about 1.0% improvement on accuracy comparing with no augmentation. This demonstrates that our generative network has a certain extrapolation ability.



## 6. Conclusion

In this paper, we propose a new CVAE-GAN model for fine-grained category image generation. The superior performance on three different datasets demonstrated its ability to generate various kinds of objects. The proposed method can support a wide variety of applications, including image generation, attribute morphing, image inpainting, and data augmentation for training better face recognition models. Our future work will explore how to generate samples of an unknown category, such as face images of a person that do not exist in the training dataset.

## References

- [1] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *NIPS 2016 Workshop on Adversarial Training*. In review for *ICLR*, volume 2016, 2017.
- [2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
- [3] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014.
- [4] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.
- [5] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems*, pages 658–666, 2016.
- [6] A. Dosovitskiy, J. Springenberg, M. Tatarchenko, and T. Brox. Learning to generate chairs, tables and cars with convolutional networks. 2016.
- [7] J. Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014, 2014.
- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
- [9] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
- [10] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.
- [11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [14] H. Larochelle and I. Murray. The neural autoregressive distribution estimator. In *AISTATS*, volume 1, page 2, 2011.
- [15] A. B. L. Larsen, S. K. Sønderby, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.
- [16] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled faces in the wild: A survey. In *Advances in Face Detection and Facial Image Analysis*, pages 189–248. Springer, 2016.
- [17] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.
- [18] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [19] V. Mnih, G. E. Hinton, et al. Generating more realistic images using gated mrf’s. In *Advances in Neural Information Processing Systems*, pages 2002–2010, 2010.
- [20] Y. Mroueh, T. Sercu, and V. Goel. Mcgan: Mean and covariance feature matching gan. *arXiv preprint arXiv:1702.08398*, 2017.
- [21] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347. IEEE, 2014.
- [22] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395, 2016.
- [23] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. *arXiv preprint arXiv:1612.00005*, 2016.
- [24] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP’08. Sixth Indian Conference on*, pages 722–729. IEEE, 2008.
- [25] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*, 2016.
- [26] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional image generation with pixelcnn decoders. *arXiv preprint arXiv:1606.05328*, 2016.
- [27] H. Permuter, J. Francos, and I. H. Jermyn. Gaussian mixture models of texture and colour for image database retrieval. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP’03). 2003 IEEE International Conference on*, volume 3, pages III–569. IEEE, 2003.
- [28] G.-J. Qi. Loss-sensitive generative adversarial networks on lipschitz densities. *arXiv preprint arXiv:1701.06264*, 2017.
- [29] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- [30] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

- [31] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.
- [32] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *AISTATS*, volume 1, page 3, 2009.
- [33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.
- [34] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.
- [35] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Motion-Based Recognition*, pages 227–243. Springer, 1997.
- [36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- [37] L. Theis, R. Hosseini, and M. Bethge. Mixtures of conditional gaussian scale mixtures applied to multiscale image representations. *PloS one*, 7(7):e39857, 2012.
- [38] L. Theis, A. v. d. Oord, and M. Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.
- [39] Z. Tu. Learning generative models via discriminative approaches. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.
- [40] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR’91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.
- [41] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.
- [42] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pages 835–851. Springer, 2016.
- [43] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
- [44] J. Xie, Y. Lu, S.-C. Zhu, and Y. N. Wu. A theory of generative convnet. *arXiv preprint arXiv:1602.03264*, 2016.
- [45] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.
- [46] L. Xu and M. I. Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.
- [47] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. *arXiv preprint arXiv:1512.00570*, 2015.
- [48] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

## Supplementary materials for: CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training

### S7. Comparing different combination of losses

In our model, we proposed to use pairwise feature matching at the image pixel level, feature level in the classification network  $C$  and the discriminative network  $D$  to update the G network. To understand the effects of each loss component, We separate the following losses:

$$\mathcal{L}_G = \frac{1}{2}(\|\mathbf{x} - \mathbf{x}'\|_2^2 + \|f_D(\mathbf{x}) - f_D(\mathbf{x}')\|_2^2 + \|f_C(\mathbf{x}) - f_C(\mathbf{x}')\|_2^2), \quad (9)$$

to three parts:

$$\mathcal{L}_G = \mathcal{L}_G(img) + \mathcal{L}_G(D) + \mathcal{L}_G(C), \quad (10)$$

where  $\mathcal{L}_G(img)$  is the  $\ell_2$  distance at the pixel level of the image,  $\mathcal{L}_G(D)$  is the  $\ell_2$  distance at the feature level in the discriminative network  $D$ ,  $\mathcal{L}_G(C)$  is the  $\ell_2$  distance at the feature level in the classification network  $C$ .

we repeated training the CVAE-GAN model with the same setting but using different combination of losses in  $\mathcal{L}_G(img)$ ,  $\mathcal{L}_G(D)$ , and  $\mathcal{L}_G(C)$ , and compared the quality of the reconstructed samples. As shown in Fig. S8, we can find that removing the adversarial loss  $\mathcal{L}_G(D)$  will cause the model to generate blurry images. Removing the pixel level reconstruction loss  $\mathcal{L}_G(img)$  caused images losing details. At last, if we remove the feature level loss  $\mathcal{L}_G(C)$  in the classification network  $C$ , the generated samples will lose category info.

On the other hand, our model produces best results. Thus, the model that we found empirically to produce realistic, diverse and category-keeping samples.

### S8. Analysis of the latent vector $\mathbf{z}$

Since the network  $G$  is able to reconstruct the input image only with the latent vector  $\mathbf{z}$  and category label  $c$ . Therefore, it is expected that it can conveniently encode all the attribute information, such as pose, color, illumination, and even more complex high level styles, in the latent vector. In this section, we will introduce some interesting findings about the latent vector.

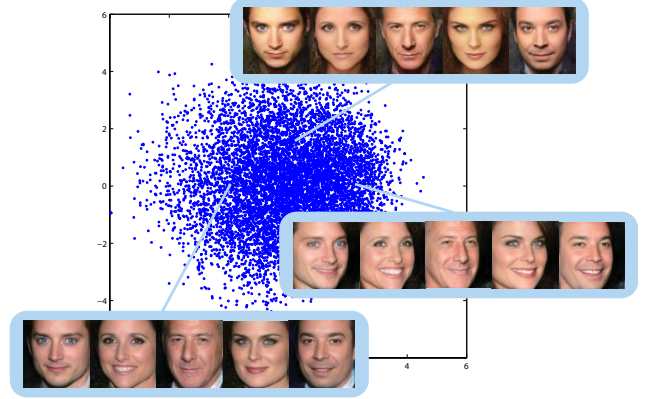


Figure S9. Illustration of the distribution of latent vector and the generated image with the same latent vector. As the readers have observed, images generated with the same latent vector  $\mathbf{z}$  but different categories will have the same attributes, such as pose, illumination, and background.

#### The same latent variable represents the same attribute.

One important finding is that, although we do not use any supervision on the attributes, the same latent vector for different labels will generate images with different category labels but with similar attributes. The reason for this phenomenon may be that images with the same attribute present certain resemblance at the pixel level. So the network automatically put them together through unsupervised clustering.

To confirm this, we train a model on a face dataset, Face-Scrub, and then extract the latent vector of all the face images. In order to clearly present the distribution, we project all the latent vectors into a two dimensional space by PCA. As shown in Figure S9, the distribution of latent vector is a Gaussian as expected, and the attribute of images, include face pose, illumination, and the background is the same for the same latent vector.

With this property, our algorithm can be used in many other applications, such as attributes transformation which generates images with different category labels but with similar attributes, and attributes retrieval which searches for other images with similar attributes.

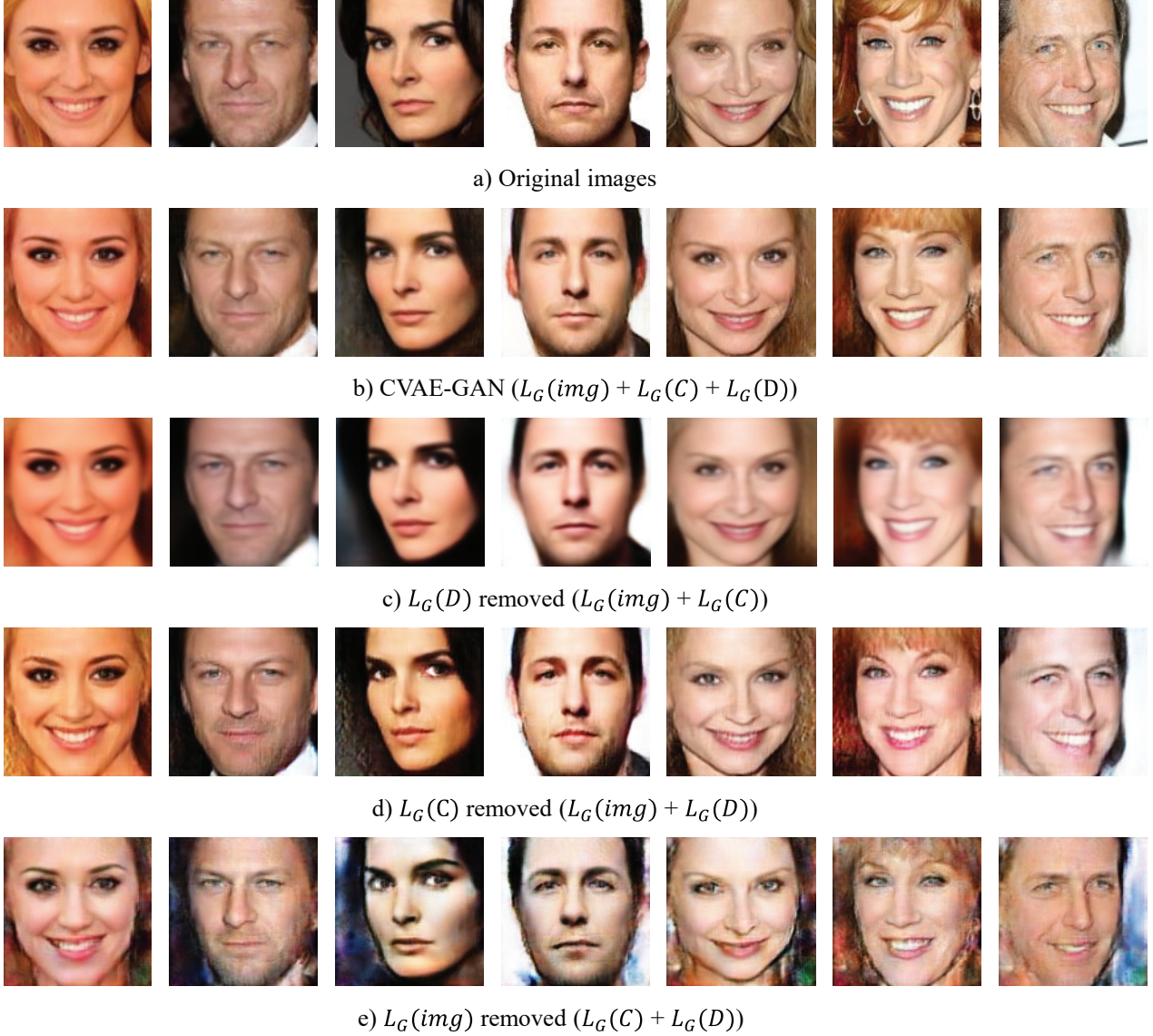


Figure S8. Visualization comparison between different generator  $G$ , each trained with different combination of losses.  $\mathcal{L}_G(img)$ ,  $\mathcal{L}_G(D)$ , and  $\mathcal{L}_G(C)$  represents pairwise  $\ell_2$  losses in pixel( $x$ ), feature in discriminative network and feature in classification network. We perform reconstruction task on each model to understand the effect of each loss in the CVAE-GAN model. a) Original face images from Facescrub datasets. b) Image reconstructed by our proposed CVAE-GAN model, you can see realistic, clear, and identity-preserving reconstruction results. c) Image reconstructed by combination of losses  $\mathcal{L}_G(img)$  and  $\mathcal{L}_G(C)$ , which shows identity-preserving but blurry results. d) Image reconstructed by combination of losses  $\mathcal{L}_G(img)$  and  $\mathcal{L}_G(D)$ , which shows realistic but sometimes identity-losing results (the 3rd and 6th image). e) Image reconstructed by combination of losses  $\mathcal{L}_G(C)$  and  $\mathcal{L}_G(D)$ , which shows reconstructed images that lose details. Comparing these reconstruction results, we come to the conclusion that  $\mathcal{L}_G(img)$  helps to preserve the detail of the image,  $\mathcal{L}_G(D)$  helps to generate sharp and clear results,  $\mathcal{L}_G(C)$  helps to keep the identity info of the reconstructed results.

### S8.1. Attributes transformation

In this part, we will experimentally validate the attributes transformation. We test our method on FaceScrub dataset. Given a source image, we first use the encoder network  $E$  to extract the latent vector  $z$ . Then using this latent vector, we can generate images in any specific category. Figure S10

shows the results of attribute transformation. We generate an image in the target category. And its attribute is similar to the source image.

### S8.2. Attributes retrieval

Through the above analysis of the latent vector, we come to such an inference: the similar latent vector represents





Figure S10. Results of attributes transformation. a) Original images from the Facescrub dataset, which offer the attributes for the generated images. b) Generated images using latent vector of the original images and the target category  $c$ . From the results, we can observe that the generated images have the same attributes as the original images.

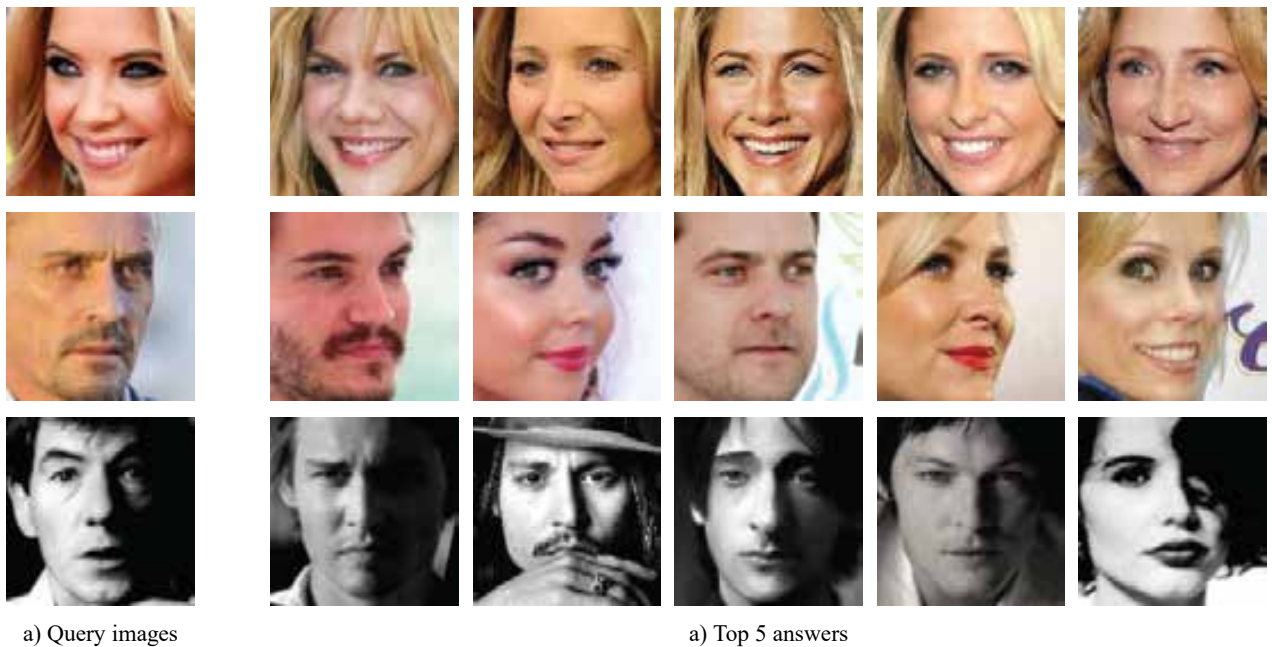
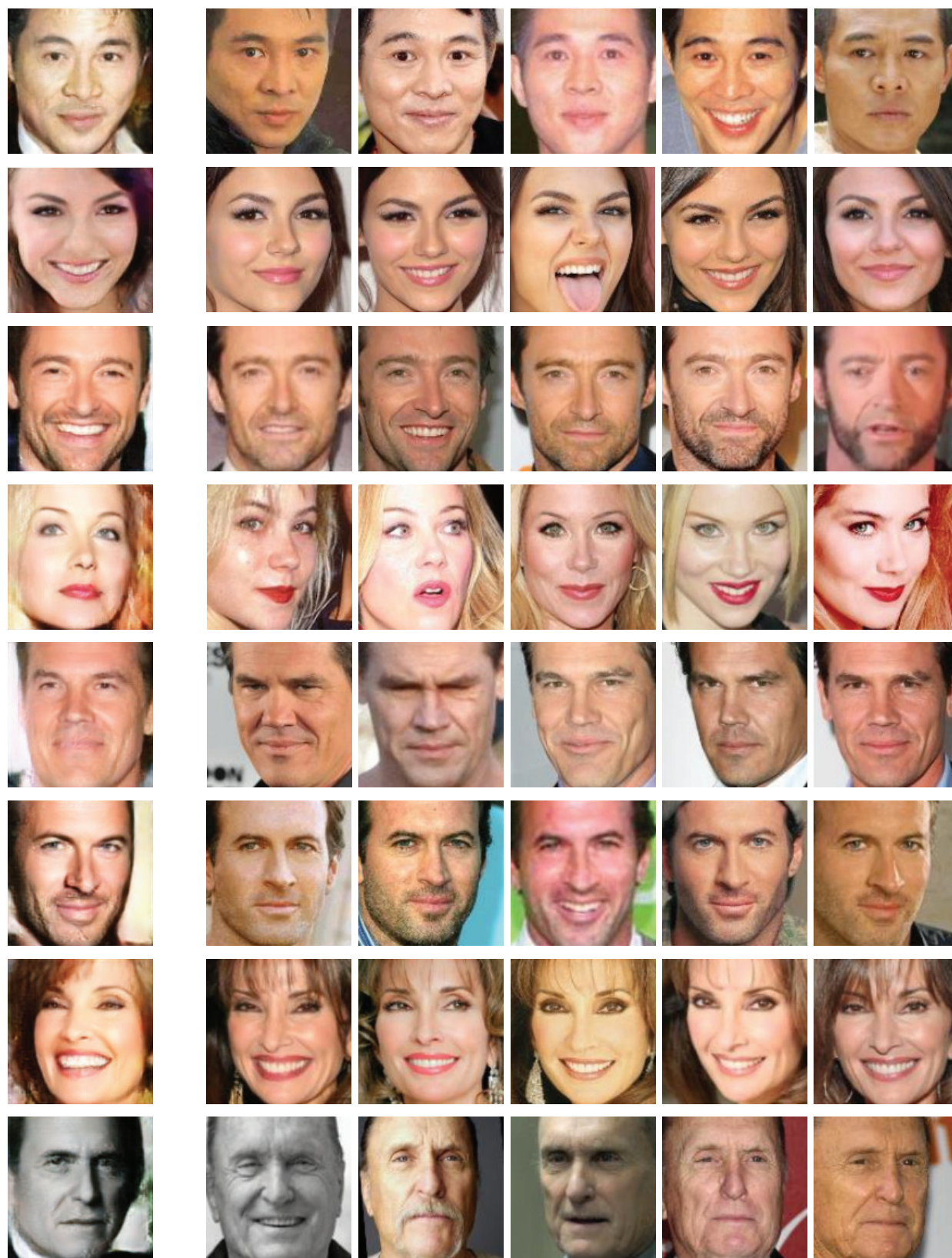


Figure S11. Results of same attributes retrieval. a) The query images. b) The top 5 answers in the original datasets.

similar attribute. Therefore, we can use our encoder model to extract the attributes features, and use them to search for the image with the most similar attribute in a dataset. We use FaceScrub dataset for this experiment. We first extracted all attribute features by the encoder network  $E$ . Then we simply conduct a image retrieval task by using  $\ell_2$  distance. As shown in Figure S11, we show the top 5 results that most similar to the query image but with different category. We found the faces with similar skin color, viewpoint or emotion.

## S9. Nearest Neighbors Test

In this section, we want to demonstrate that our model is not just memorizing all training examples in the training process. Our model can generate samples which are not copies of the training samples. We choose FaceScrub dataset for this experiment. Firstly, we randomly generate 8 samples from 8 different categories. Then we simply conduct an image retrieval task using  $\ell_2$  distance at the pixel level. As shown in Fig. S12, we show the top 5 results that most similar to the generated images, we find that the generated samples have the same identity but different attributes like pose, illumination, and emotion as compared to nearest neighbors.



a) Generated images

b) Top-5 nearest neighbors in the origin datasets

Figure S12. Results of nearest neighbors search on the generated samples. a) The generated samples from the CVAE-GAN model. b) The top-5 nearest neighbors real samples, which shows the novelty of generated images.